IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models

Abstract

Recent years have witnessed the strong power of large text-to-image diffusion models for the impressive generative capability to create high-fidelity images. However, it is very tricky to generate desired images using only text prompt as it often involves complex prompt engineering. An alternative to text prompt is image prompt, as the saying goes:"an image is worth a thousand words". Although existing methods of direct fine-tuning from pretrained models are effective, they require large computing resources and are not compatible with other base models, text prompt, and structural controls. In this paper, we present IP-Adapter, an effective and lightweight adapter to achieve image prompt capability for the pretrained text-to-image diffusion models. The key design of our IP-Adapter is decoupled cross-attention mechanism that separates cross-attention layers for text features and image features. Despite the simplicity of our method, an IP-Adapter with only 22M parameters can achieve comparable or even better performance to a fully fine-tuned image prompt model. As we freeze the pretrained diffusion model, the proposed IP-Adapter can be generalized not only to other custom models fine-tuned from the same base model, but also to controllable generation using existing controllable tools. With the benefit of the decoupled cross-attention strategy, the image prompt can also work well with the text prompt to achieve multimodal image generation. The project page is available at \url{https://ip-adapter.github.io}.